Higher Fidelity Subtree Merging for Disk-Covering Methods

Nicholas D. Pattengale[1], Krister M. Swenson[1,2], Monique M. Morin[1], Bernard M.E. Moret[1,2,3]
[1]*Department of Computer Science, University of New Mexico, USA*
[2]*Laboratory for Computational Biology and Bioinformatics,*
School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland
[3] *Swiss Institute of Bioinformatics, Lausanne, Switzerland*

Disk-Covering methods (DCMs) are a family of divide and conquer strategies for large-scale phylogeny reconstruction. Merging subproblems, as in any divide and conquer strategy, is a critical aspect of all DCMs. In order to avoid introducing unnecessary error, the traditional DCM subtree merging technique is highly conservative. We present a framework for better discerning troublesome components in subproblems. We in turn utilize our framework as a basis for a new subtree merging algorithm.

I. INTRODUCTION

Most problem formulations for phylogenetic reconstruction are computationally hard[1, 2]. Approximation algorithms are generally ineffective in this regime, as often solutions must be within orders of magnitude less than one percent from optimal in order to be useful[3]. The most effective approaches, in practice, tend to be heuristics that do not reasonably scale to more than a few thousand taxa. As systematic biologists desire to reconstruct high-quality phylogenies for datasets of tens to hundreds of thousands of taxa, methods must be pioneered in order to accommodate such large datasets. Disk-covering methods constitute a family of algorithms that address large datasets.

II. TERMINOLOGY

A *phylogenetic tree* T is an undirected, leaf-labeled, connected, acyclic graph. The leaves of the tree, also called *taxa* or *tips*, form a set S where $|S| = n$. Removing an edge e from T breaks the tree into two smaller trees – T_r and T_l (with leaf sets S_r and S_l, respectively). Thus the effect of e is to *split* S into $S_r|S_l$. A split is *trivial* if either $|S_r| = 1$ or $|S_l| = 1$. The set of non-trivial splits induced by all of the edges in T is denoted $\Sigma(T)$. It is the case that $|\Sigma(T)| \leq (n-3)$. The inequality becomes an equality whenever the tree is binary (i.e. contains no *polytomies* – internal nodes of degree greater than three), which is also referred to as *fully resolved*. A *consensus method* (see survey in [4]) is an algorithm that takes as input a set of trees (all with the same taxa) and returns a single "summary" tree (over the same set of taxa).

III. DISK-COVERING METHODS (DCMs)

DCMs[5–8] conceptually resemble a standard divide and conquer approach. One notable deviation from standard divide and conquer is that subproblems must be overlapping in order to enable merging. DCMs proceed by decomposing the set of taxa into overlapping subsets. Once suitably decomposed, the subproblems are solved using a standard phylogenetic reconstruction algorithm, i.e. TNT, PAUP*, RaxML, GARLI, POY, *et. cetera*. Finally the subproblems are reassembled into a single tree. In certain DCM variants (notably, Rec-I-DCM3[7]) the entire process is repeated numerous times in pursuit of iterative refinement. Another DCM (DCM-GRAPPA[8]) addresses large-scale reconstruction with gene orders instead of sequence data.

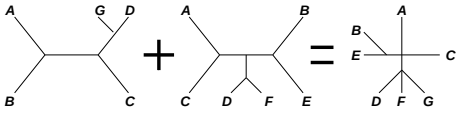

FIG. 1: The strict consensus merger often yields polytomies.

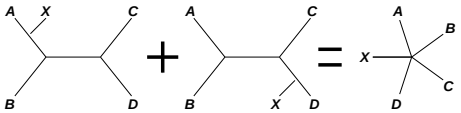

FIG. 2: Removing the taxon X would uncover structural similarity.

To date, DCMs have handled subproblem merging with the so-called *Strict Consensus Subtree Merger (SCM)*, a conservative approach that discards all edges (in the subproblem overlap) that are not common to all subproblems. Fig. 1 illustrates a common occurrence with SCM – the introduction of polytomies. DCMs resolve polytomies by random refinement and as such inevitably give rise to unnecessary error.

IV. ROGUE TAXA

We are interested in detecting situations when removing a small subset of taxa uncovers strong structural similarities between the induced subtrees. In other words, is it possible to avoid some of the polytomies arising from SCM? Fig. 2 illustrates a case in which removing a single taxon reveals strong structural similarities among the induced subtrees.

In the preceding case the removal of a *single* taxon dramatically improved consensus resolution. There are, however, more subtle situations for which it would be desirable to detect. In fact there are cases in which removing a small subset of taxa U produces results nearly as dramatic as the preceding example, yet removing any subset $V \subset U$ does not improve resolution whatsoever. Fig. 3 illustrates such a case. Obviously, there is a tradeoff between increasing resolution and removing so many taxa so as to discard valuable information.

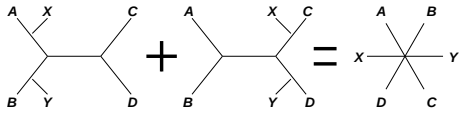

FIG. 3: Removing any subset of $U = \{X, Y\}$ does nothing to improve resolution, yet removing U itself dramatically improves resolution.

V. FORMALIZING TAXON IMPACT

Since we wish for our framework to be applicable in the context of any consensus method, we refer in a general fashion to consensus method C. Thus for a set of trees $\bigcup_i T_i$ (over taxa S) and a consensus method C we define the *impact* of a subset of the taxa $U \subset S$ informally as the improvement in resolution between the consensus of the original tree set and the consensus of the induced subtrees by removing U from each. Formally,

$$I(U) = |\Sigma(C(\cup_i(T_i - U)))| - |\Sigma(C(\cup_i T_i))| + r - \sum_{V \subset U} I(V)$$

where r is a corrective factor to account for non-trivial bipartitions in $C(\bigcup_i T_i)$ becoming trivial when removing U. It may be desirable to incorporate a penalty proportional to $|U|$ in order to accommodate the concern of losing too much information. It may also be desirable to introduce a normalization factor.

VI. REMOVABLE TAXA CONSENSUS MERGER

As the expression for taxa impact is defined in terms of subsets of U, scoring the impact of any taxa subset of size $\Theta(n)$ implies the examination of an exponential number of subsets. While we have not yet attempted to prove so, we suspect that most natural optimization problems based upon the full definition of taxa impact will be $\mathcal{NP}$-hard. We remark, however, that restricting the definition of impact to all subsets of fixed size k yields a computation that scales no worse than $O(mn^k \cdot g(n, m))$, where $g(n, m)$ indicates the complexity of the employed consensus method. Thus it is reasonable to expect acceptable performance for a merge routine examining all subsets of size $k \leq 3$ (give or take, based on the situation). This small value should be sufficiently small in practice (most DCMs bound subproblem size, so n should be relatively small), as well as address the concern of removing too many taxa in the name of resolution.

If the $k \leq 3$ constraint is worrisome (if, for example, the concept of taxa impact were utilized in a setting other than DCM merge), another possibility is to examine all subtrees. For a set of m trees over n taxa there are no greater than $O(mn\log n)$ such subtrees. It is also the case that the sum of the sizes of all subtrees is also $O(mn\log n)$ and thus examining all subtrees would constitute a perfectly reasonable, and tractable, approach. It is natural to conjecture that this approach is relevant in situations where the reconstruction method is based upon pruning and regrafting subtrees (clades), and is prone to misplacing whole subtrees.

We finally present a new subtree merger based upon taxa impact. Assuming a threshold parameter p and a constant k, calculate the taxa impact for subsets of size less than or equal to k and of all subtrees. Whenever subsets are identified whose impact exceeds p, exclude the subset from the source trees, and then apply the strict consensus merger as before. In an application regime such as Rec-I-DCM3 the implication is that the reassembled (and subsequently decomposed) tree is missing taxa. However, our aim is that by guiding the search with strong components, we will better accomodate reincorporating rogue taxa.

VII. FUTURE WORK

There are many ways to define taxa impact. In this extended abstract we have presented one definition, however we plan to investigate others. In addition, we plan to implement and perform experimental evaluation of the proposed subtree merger routine in the context of a new disk-covering method.

[1] Roch, S.: A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. IEEE/ACM Transactions on Computational Biology and Bioinformatics **03** (2006) 92–94

[2] Foulds, L., Graham, R.: The steiner problem in phylogeny is np-complete. Adv. Appl. Math. (1982) 43–49

[3] Williams, T.L., Berger-Wolf, T., Moret, B.M., Roshan, U., Warnow, T.: The relationship between maximum parsimony scores and phylogenetic tree topologies. In: U. New Mexico TR-CS-2004-04. (2004)

[4] Bryant, D.: A classification of consensus methods for phylogenies. In Janowitz, M., LaPointe, F.J., McMorris, F., Mirkin, B., Roberts, F., eds.: Bioconsensus. Volume 61 of DIMACS Series., American Mathematical Society, Providence (2003) 163–184

[5] Huson, D.H., Nettles, S., Warnow, T.: Disk-covering, a fast-converging method for phylogenetic tree reconstruction. Journal of Computational Biology **6** (1999)

[6] Huson, D.H., Vawter, L., Warnow, T.: Solving large scale phylogenetic problems using dcm2. In Lengauer, T., Schneider, R., Bork, P., Brutlag, D.L., Glasgow, J.I., Mewes, H.W., Zimmer, R., eds.: ISMB, AAAI (1999) 118–129

[7] Roshan, U., Moret, B.M.E., Warnow, T., Williams, T.L.: Rec-i-dcm3: A fast algorithmic technique for reconstructing large phylogenetic trees. In: CSB, IEEE Computer Society (2004) 98–109

[8] Tang, J., Moret, B.M.E.: Scaling up accurate phylogenetic reconstruction from gene-order data. In: ISMB (Supplement of Bioinformatics). (2003) 305–312